COMPARING THREE METHODS TO CREATE MULTILINGUAL PHONE MODELS FOR VOCABULARY INDEPENDENT SPEECH RECOGNITION TASKS

Joachim Köhler

German National Research Center for Information Technology (GMD)
Institute for Media Communication (IMK),
53754 Sankt Augustin, Germany
Joachim.Koehler@gmd.de

ABSTRACT

This paper presents three different methods to develop multilingual phone models for flexible speech recognition tasks. The main goal of our investigations is to find multilingual speech units which work equally well in many languages. With this universal set it is possible to build speech recognition systems for a variety of languages. One advantage of this approach is to share acoustic-phonetic parameters in a HMM based speech recognition system. The multilingual approach starts with the phone set of six languages ending up with 232 language-dependent and context-independent phone models. Then, we developed three different methods to map the language-dependent models to a multilingual phone set. The first method is a direct mapping to the phone set of the International Phonetic Association (IPA). In the second approach we apply an automatic clustering algorithm for the phone models. The third method exploits the similarities of single mixture components of the language-dependent models. Like the first method the language specific models are mapped to the IPA inventory. In the second step an agglomerative clustering is performed on density level to find regions of similarities between the phone models of different languages. The experiments carried out with the SpeechDat(M) database show that the third method yields in almost the same recognition rate as with language-dependent models. However, using this method we observe a huge reduction of the number of densities in the multilingual system.

1. INTRODUCTION

Over the last years automatic speech recognition systems have reached a level of quality which allows the introduction of commercial products. However, a new problem has occurred: the language-dependency of current recognition technology. The phonetic models used in state-of-the-art systems are extremely language-dependent. The overall goal of our research activities is to create a multilingual and almost language independent recognition system which works in the most important languages of the world. We started our multilingual approach with OGI MLTS database [15] based on the work of [1]. Nowadays, even larger multilingual databases are available like SpeechDat(M)[1], Call-Home etc. These databases allow a robust modeling of phonetic units for different languages. Instead of using language-dependent acoustic models our approach tries to exploit the acoustic-phonetic similarities of sounds across languages. This approach has two main advantages. First, the number of HMM parameters can be reduced significantly if it is possible to share phone models in different languages. Second, these multilingual models speed up the process of cross-language transfer. With the multilingual phone models the huge data collection process can be avoided or at least it can be reduced. This paper shows different approaches to achieve the goal to exploit the acoustic-phonetic similarities.

The paper is organized as follows: First, we present three different methods to create multilingual phone models using HMM technology. Then we perform our experiments with a language-dependent system covering six languages. These multilingual experiments are then given in the following chapter. At the end we give a summary of the current research status and a perspective for future research activities.

[1]For information about SpeechDat see the following URL's:
http://www.phonetik.uni-muenchen.de/SpeechDat.html
http://www.icp.grenet.fr/ELRA/home.html

2. MULTILINGUAL PHONE MODELING

This section shows different approaches to find multilingual phone models for automatic speech recognition tasks. One central problem is to detect and to exploit the acoustic-phonetic similarities across languages. Which sound in one language is similar enough to a sound of another language to provide only one common model? This question leads to the definition of a similarity measurement of speech sounds. The other question is, if the phone is the optimal entity to exploit the similarities. Or is another speech unit like a sub phone unit or a single density of a continuous density HMM (CDMM) more appropriate to create multilingual models. The overall goal of the different approaches to find multilingual speech units is to generate models which perform as well as language-dependent models for different recognition tasks. Thus, we can formulate the task to create accurate acoustic models which also exploit the similarities across languages.

2.1. Mapping to the IPA based phone set (IPA-MAP)

The most obvious approach is to map the language-dependent models to the appropriate phone of the inventory of the International Phonetic Association (IPA). Here, the phonetic mapping is performed with phonetic knowledge rather than with some statistical based similarity measurement. Most of the phonetic inventories which are in use are based on IPA, like SAMPA, WORLDBET, TIMITBET or SPICOS. The rule of the mapping of the language-dependent phones $Ph_{l,i}^{LDP}$ to the multilingual phone units is:

$$Ph_{l,i}^{LDP} \rightarrow Ph_{j}^{IPA} \quad (1)$$

The mapping is performed for each language. All phonetic segmentation and transcription files (label files) are transformed to

the IPA based inventory. After this mapping a Viterbi based HMM Maximum Likelihood training is performed. Figure 1 shows the different steps of the approach IPA-MAP.

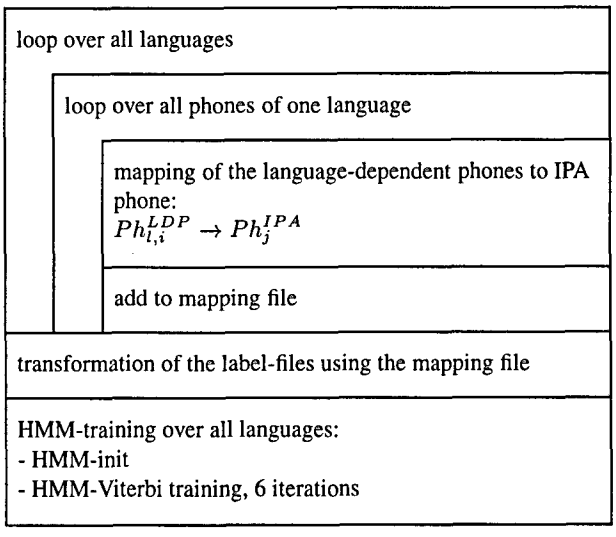

Figure 1: Algorithm IPA-MAP

The main advantage of this approach is the simple way of getting multilingual models. Further, the final IPA-based models have a clear representation in the multilingual context and the cross-language transfer is also very simple. The sounds of the new language can be extracted very easily from the multilingual phone library. On the other hand the direct use of IPA does not consider the spectral properties and the statistical similarities of the phone models. Further, the IPA-based units do not model some language-dependent properties of the sounds. This can yield in a decrease of the accuracy of the acoustic models. This problem will be more severe as more languages will be included in this approach. Another disadvantage is that some inconsistencies of different phone systems of different languages and inventories can hurt this method.

2.2. Multilingual Phone Clustering (MUL-CLUS)

In this approach the language-dependent phone models are mapped to a multilingual set using a bottom-up cluster algorithm. Therefore, a similarity between two phone models has to be defined. In this work we apply a log-likelihood LL based distance measure. The distance between two phone models λ_i and λ_j is:

$$D_{LL}(\lambda_i, \lambda_j) = LL_i^i - LL_j^i \quad (2)$$

$$D_{LL}(\lambda_i, \lambda_j) = \log p(X_i|\lambda_i) - \log p(X_i|\lambda_j) \quad (3)$$

where λ_i is the model of phone i. The data is given by the token X_i. Respectively, the distance $D_{LL}(\lambda_j, \lambda_i)$ is given by:

$$D_{LL}(\lambda_j, \lambda_i) = LL_j^j - LL_i^j \quad (4)$$

$$D_{LL}(\lambda_j, \lambda_i) = \log p(X_j|\lambda_j) - \log p(X_j|\lambda_i) \quad (5)$$

Because the distances are not symmetric we calculate the average distance:

$$D_{LL}(\lambda_i; \lambda_j) = \frac{1}{2}(D_{LL}(\lambda_i, \lambda_j) + D_{LL}(\lambda_j, \lambda_i)) \quad (6)$$

At each cluster step the most similar pair of clusters are merged to a new cluster. This means that the two clusters $\hat{C}_i$ and $\hat{C}_j$ of all cluster pairs C_i and C_j with the smallest distance are merged:

$$(\hat{C}_i, \hat{C}_j) = \underset{C_i, C_j}{\operatorname{argmin}} D(i, j) \quad (7)$$

Because the estimation of the new phone models of the merged cluster is difficult to achieve the distance is always computed with the original language-dependent models which are the basic elements of one cluster. Hence, the distance between two clusters are determined with the furthest neighbor criterion. Therefore, we calculate the maximum distance of the initial clusters C_k^0 and C_l^0 which are in this case the language-dependent phone units.

$$(\hat{C}_i, \hat{C}_j) = \underset{k \in C_i, l \in C_j}{\operatorname{argmax}} D(k, l) \quad (8)$$

The usage of the furthest neighbor criterion has also the advantage to avoid huge log-likelihood calculations. The calculation of equation 6 requires also the data of the phone models. The data corresponds to the phone tokens which are extracted from the phonetic label files. Each phone has a pool of tokens which are used for the distance calculation. The number of tokens of each language-dependent phone unit is set to 500.

The complete algorithm to create multilingual phone models using clustering methods is given in figure 2.

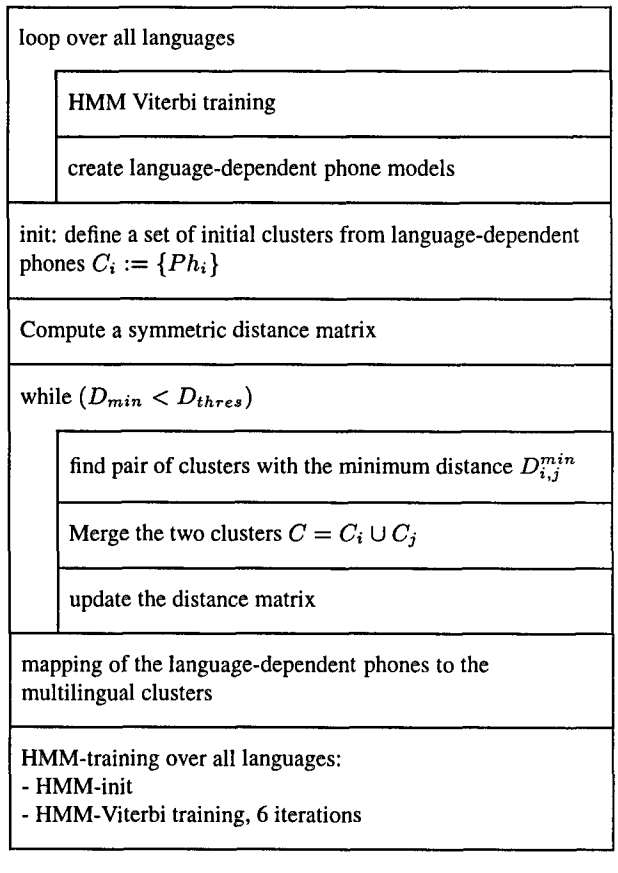

Figure 2: Algorithm to create multilingual phone models using phone distance measurement and clustering (MUL-CLUS)

The cluster process continues until all calculated cluster distances are higher than a pre defined distance threshold. Alternatively, the clustering stops if a specified number of final clusters is achieved. After the clustering is finished we can use the cluster information to map the language-dependent models to the multilingual inventory. All label files are processed with this mapping information. Then the HMM models are trained with the maximum likelihood based Viterbi training.

The automatic clustering has the advantage to use statistic measurement based on HMM technology which is also used during recognition. The disadvantage is that the final multilingual units lose some clear representation and it is more difficult to transfer this models to a new language.

2.3. IPA-based Density Clustering (IPA-OVL)

The previous two approaches try to create complete multilingual phone models. This means that all parameters (i.e. sub phone units, densities of a CDHMM) of one model are shared across the different languages. On the other hand there are several language specific properties of the sounds. They exist due to different phonetic context, speaking style and rate, prosodic features and allophonic variations. To cover these effects we have presented a novel approach to create multilingual phone models [15]. Instead of complete overlapping phone models we assume that there are language-independent realization. This approach is achieved by using mixture densities. Figure 3 shows the idea of this method. There are regions of one IPA sound which are used in one, two or three languages. In this example the nasal

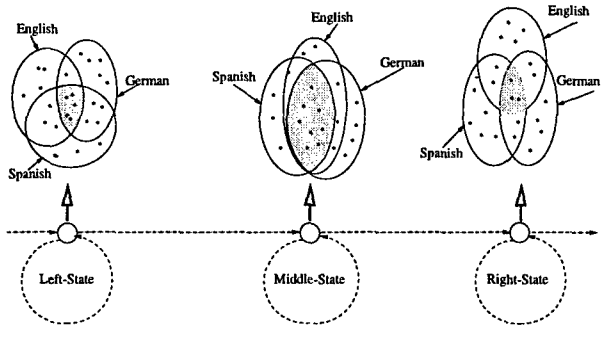

Figure 3: principles of the method IPA-OVL (two dimensional case).

[m] occuring in the languages German, Spanish and English has mixture components which are used in one, two or all three languages.

The creation of the multilingual models is shown in figure 4. First, the language-dependent models are trained as before. Each language-dependent phone consists of 3 segments (sub phone units) each modeled by a mixture density. This is expressed by:

$$\lambda_{l,p}^{mono} = \left\{ S_{l,p,1}^{mono}, S_{l,p,2}^{mono}, S_{l,p,3}^{mono} \right\} \qquad (9)$$

where l is the language index and p in the phone index.

In the second step the mixtures of the language-dependent segments which belong to the same IPA-based phone are collected in one common pool of densities. Then we apply an hierarchical agglomerativ cluster algorithm to find and merge similar densities. The clustering is performed for each segment separately.

Because we work in our system with global variance values we use only the mean vectors for clustering. As distance measure giving the similarity between μ_i und μ_j the weighed L1-norm is applied:

$$D(\lambda_j;\lambda_i) = \frac{N_j N_i}{N_j + N_i} \sum_{d=1}^{D} |\mu_{i,d} - \mu_{j,d}| \qquad (10)$$

In previous investigations we found that is important to normalize the distance by the number of occurrences N_i und N_j which give information how often the densities are seen during training. This normalization avoids the generation of very big clusters which dominate the small clusters. One important aspect is that all clusters should have a similar number of elements. Otherwise the resulting clusters lose their power to discriminate between different sounds.

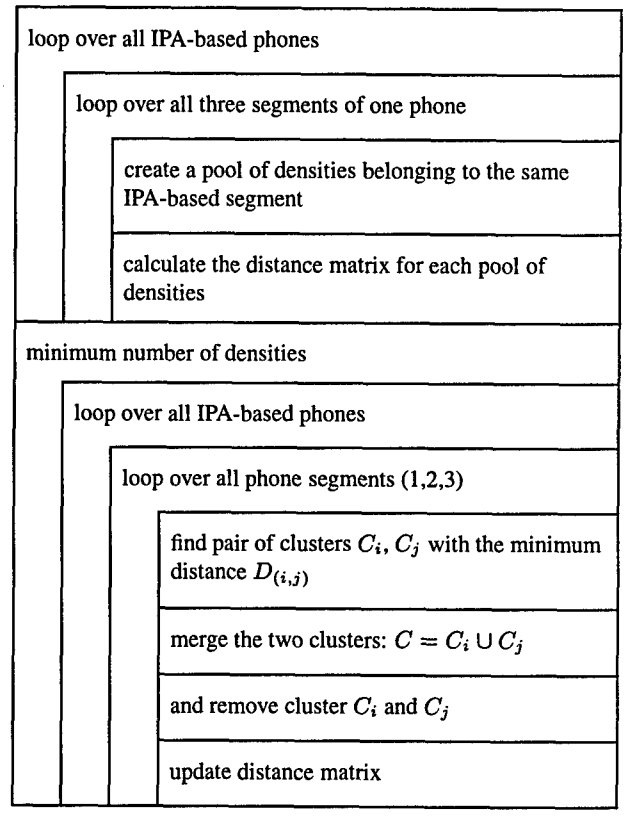

Figure 4: Algorithm to create multilingual mixture densities (IPA-OVL)

For each pool of densities a distance matrix is calculated using equation 10. After each clustering step the overall number of densities is reduced by one element. The new cluster is given by the averaged mean vector of the two merged clusters. The clustering is finished if the complete system has a pre-defined number of densities. After finishing the cluster algorithm we have for each IPA-based phone a multilingual mixture density. Whereas the mixture density has multilingual regions the mixture weights are still language-dependent. For the calculation of the emission probabilities we use:

$$b_s(\vec{x}) = \sum_{m=1}^{M_s} c_{s,m}^{LDP} \mathcal{N}(\vec{x}, \vec{\mu}_{s,m}^{IPA}) \qquad (11)$$

Hence, this approach has some similarities to the semi-continuous HMMS. However, here the densities are shared only for one segment of one IPA-based phone across different languages. As final step the parameters of the multilingual mixture densities are reestimated during a Viterbi training. With this kind of multilingual modeling we also achieve a huge reduction of parameters in multilingual system. The combination of language-specific properties and automatically detection of multilingual realizations we exploit the acoustic-phonetic similarities in an optimal way.

3. EXPERIMENTS

In this section we perform several tests to compare the multilingual approaches. First, we describe briefly the speech engine. Second, we present the multilingual system using the language-dependent models. This system serves as comparison to the three previously described methods.

3.1. Description of the HMM-Based ASR system

For our investigations we use the SIEMENS HMM-based speech engine. The feature extraction generates every 10 ms a frame consisting of 24 mel-scaled cepstral, 12 Δ cepstral, 12 $\Delta\Delta$cepstral, 1 energy, 1 Δ energy and 1 $\Delta\Delta$ energy components. Each frame is processed by a LDA transformation reducing the 51 components to 24 values. To work in a multilingual environment one single LDA is calculated for all different languages. The acoustic models are based on Continuous Density HMMs (CDHMM) with Gaussian density functions. In our investigations we work only with context-independent models which consist of 3 sub-phone units (phone segments). Each segment is modeled by two states with tied emission probability.

3.2. Multilingual System with language-dependent models

The multilingual system covers the six languages American English, French, German, Italian, Portuguese and Spanish. The speech material is taken from the SpeechDat(M) and the Macrophone databases. Because all databases have only an orthographic transcription, all systems must be bootstrapped to generate an initial segmentation and label files. The bootstrapping was carried out with multilingual phone models based on the IPA-MAP method. The evaluation and tests were carried out on word and phone level. The word recognition rates are important for a final application and the phone recognition rates give some detail information about the acoustic modeling accuracy.

The training of the models is performed with the phonetic rich sentences of the databases. This should guarantee the vocabulary independence of the acoustic models. These models are also called *Type-In* models. The amount and structure of the training and test material is given in table 1. The training is performed with more than 4000 speakers and more than 35K sentences. The duration of the training material is almost 32 hours of pure speech without silence. The overall number of language-dependent phone units is 232. Italian has the greatest number of phones (49) because the SAMPA inventory distinguish between short and long consonants. Spanish has the smallest number using only 31 phones. The complete system has 31999 densities which means that in average each of the 232 language-dependent phone models have 45 densities.

After the training the models are tested on an isolated word and a phone recognition task. The recognition results for isolated words are summarized in table 2. The vocabulary size of this task varies between 47 and 70 words for the languages taken from SpeechDat(M). For American English the vocabulary size is 685 because there is no core test set for application words. The best results are achieved for German (96.6%). Also for the other 4 European languages we get results better than 90%. The result for American English is only 64.9% due to the high perplexity of the recognition task.

	#speaker tr-dev-te	#utt. Tr-Utt	hour.min Tr-Time	# phones
French	667-166-167	6.0K	5.03	37
German	667-166-167	5.0K	4.18	38
Italian	667-166-167	5.8K	4.15	49
Portuguese	667-166-167	5.9K	7.33	38
Spanish	667-166-167	6.0K	5.38	31
Am.-English	1000-500-500	6.4K	5.12	39
Overall	4335-1330-1335	35.1K	31.59	232

Table 1: Structure of the training and test databases using SpeechDat(M) and Macrophone: tr $\hat{=}$ number of speakers for training; dev $\hat{=}$ number of speakers for developing purposes; te $\hat{=}$ number of speakers for testing; Tr-Utt $\hat{=}$ number of phonetic rich training sentences; Tr-Time $\hat{=}$ time and duration of phonetic rich training sentences; number of phone units per each language

Language	#Rec-. Tokens	Voc. Size	*Rec.-Rate*
French	1420	57	92.2%
German	949	49	96.6%
Italian	983	47	94.4%
Portuguese	931	61	93.0%
Spanish	1242	70	93.3%
Am.-English	2612	685	64.9%
Average	–	–	89.0%

Table 2: Isolated word recognition rate for SpeechDat(M) and Macrophone database; Rec-Tokens: number of tested words; Voc. size: size of the vocabulary (perplexity); Rec. rate: word recognition rate

In the second test phone recognition rates are measured. The results given in table 3 including insertions, deletions and substitutions. For the continuous phone recognition task language-dependent bigram models are used to achieve a higher phone accuracy. It is very obvious that for Spanish and Italian the best phone recognition rates are achieved (56.9% and 53.2%). Both languages have a clear vowel structure. Also for German, French and Portuguese the recognition rates varies between 47.0% and 48.5%. Only for American English the recognition result ends

Language	#Rec-. Tokens	Voc. Size	*Phone Acc.*
French	12964	37	48.3%
German	12839	38	48.5%
Italian	10804	49	53.2%
Portuguese	21751	38	47.0%
Spanish	17512	31	56.9%
Am.-Englich	10815	39	37.7%
Average	–	–	48.6%

Table 3: Continuous phone recognition rate for SpeechDat(M) and Macrophone including deletions, insertions and substitutions

	LDP	IPA-MAP	MUL-CLUS	IPA-OVL
French	92.2%	90.9%	90.8%	92.5%
German	96.6%	91.6%	94.8%	96.5%
Italian	94.4%	93.6%	94.0%	93.7%
Portuguese	93.0%	89.6%	91.9%	91.9%
Spanish	93.3%	92.5%	93.3%	93.1%
Am.-English	64.9%	56.5%	57.0%	63.2%
Average	89.0%	85.5%	86.9%	88.5%

Table 4: isolated word recognition rates using the different multilingual approaches

with a disappointing 37.7% rate. One reason for this result could be the quality of the orthographic and phonetic transcription of the Macrophone database. In other investigation the results for American English are very similar to results in French or German [18].

Altogether the results on word and phone level show that it is possible to create task independent models with phonetic rich training material. These models are compared in the following section with the multilingual approaches.

3.3. Results using the Multilingual Approaches

Table 4 summarizes the isolated word recognition rates of the three different approaches in comparison to the language-dependent modeling. For these tests the number of densities was almost the same to achieve a fair comparison. The method IPA-OVL outperforms the other two methods (IPA-MAP and MULS-CLUS) and it was nearly as good as with the language-dependent models. The decrease in recognition rate was only 0.5% with only 13K densities instead of 31K densities in the language-dependent case. Hence, the method IPA-OVL is able to detect and exploit the acoustic-phonetic similarities across the phones of different languages. The data-driven phone clustering approach (MUL-CLUS) performs also better than the direct and simple mapping to the IPA inventory. For this two methods which model complete multilingual phones the decrease of recognition rate was 3.5% (IPA-MAP) and 2.1% (MUL-CLUS). Before we give a final conclusion the detailed results of the three methods are discussed.

IPA-MAP

The method IPA-MAP maps the 232 language-dependent models to 95 multilingual models. There are 13 phones (plosives, fricatives and nasals) which occur in all six languages. Table 5 gives an overview how many phones are used in different languages. This table also shows that 48 phones are still monolingual because they occur only in one language. However, the number of system parameters is drastically reduced. The number of densities decreases from 31999 to 13555 which reduces memory and computational resources of the multilingual recognition system significantly. However, the isolated word recognition rate decreases from 89% to 85.5%.

Whereas the decrease for the four Romance languages is small the reduction for German and American English is 5.0% and 8.4% respectively. Possible explanations for this effect are:

- differences in the quality and recording conditions of Macrophone and SpeechDat(M) databases:
  Although a channel compensation algorithm is used not all differences in the databases can be removed. This would at least explain the reduction of the American system.

# La.	# Ph.	list of phones
6	13	b d f g j k l m n p s t z
5	7	ʃ ɔ a r u v w
4	7	ɛ ŋ ɲ ʒ e i o
3	3	ə ʎ tʃ
2	17	ʊ œ ɽ ʁ ɑ ʏ ɛ̃ ɔ̃ ĩ aɪ aʊ dʒ h iː sː x θ
1	48	æ ç ɪ ø øː β ɛː ʃː ɣ ðː ɔʏ ɜ ʎː ɐ ɳ ʌ ɑ̃ œ̃ ĩ õe aː bː dː dʒː dz eː ei fː gː jː j̃ kː lː mː nː oː oʊ pː pf tʃː tː ts uː ũ vː w̃ y ɔi

Table 5: Multilingual inventory using IPA-MAP

- sensitivity of the models for big vocabulary size:
  If the recognition task has a very high perplexity (in this case it is 685) very exact acoustic models are required. A small degradation of the models yields in a severe reduction of recognition rate.
- dominance of the Romance language in comparison to Germanic languages:
  Four of the six languages belong to the Romance language family. Hence, the multilingual models are dominated by the Romanian languages. This would explain the decrease of the German system.
- Inconsistency of the different phone inventories:
  Whereas for the Romance languages SAMPA is used, the German lexicon is based on SPICOS and the American lexicon uses TIMITBET. Although all inventories tries to realize the IPA-inventory there are some inconsistencies and problems during the mapping. For example in SPICOS the affricates [tS], [dZ], [pf] and [ts] are divided in two single phones. Also in the CMU-lexicon we observed some differences to the other inventories which could not be resolved easily. The central phone [ɐ] and the back vowel [ʌ] have the same phoneme symbol / ah /. Hence, the same symbol / ah / is used to transcribe the words "bottom" / b aa t **ah** m / and "cut" / k **ah** t /.

MUL-CLUS

The data-driven method MUL-CLUS yields in a higher recognition rate than the method IPA-MAP. Especially for German the results are much better. Instead of a reduction of 5.0% we observe only a decrease of 1.8%. However, the reduction for American English is still very obvious (7.9%). For this experiment the final number of multilingual phone units was chosen to 95 to have the same number of phones as before. The remaining clusters differs from the IPA-based mapping. The biggest cluster contains the fricatives [f], [s] of all six languages. Table 6 shows a selection of generated phone clusters. There are also some clusters which have same elements as with the IPA-MAP method. These clusters contain the nasals [m] and [n]. Phones which differ only in the phonetic length are very often mapped to the same cluster, especially for consonants. However, we also have 50 clusters with only one element. This means that we have still a huge number of monophones. Further, experiments were carried out with a varying size of final multilingual phone clusters. An observable decrease in recognition rate was observed when the 232 language-dependent models were clustered to less than 130 multilingual phones.

IPA-OVL

Here the clustering was performed on density level. The final

#CL	Cluster elements
15	f^{AE} f^{SP} f^{IT} f^{GE} f^{PT} f^{FR} f:^{IT} s^{AE} s^{GE} s^{PT} s^{FR} s:^{IT} s^{SP} s^{IT} θ^{SP}
12	p^{AE} p^{SP} p^{IT} p^{FR} p^{PT} p^{GE} t^{SP} t^{IT} t^{PT} t^{FR} t^{GE} t:^{IT}
10	j^{AE} i:^{AE} i^{SP} i^{IT} i^{PT} i^{FR} i:^{GE} j^{SP} j^{IT} j^{PT}
7	m^{AE} m^{SP} m^{IT} m^{FR} m^{PT} m^{GE} m:^{IT}
7	n^{AE} n^{SP} n^{IT} n^{GE} n^{FR} n^{PT} n:^{IT}

Table 6: Selection of multilingual phone clusters generated with MUL-CLUS

number of densities was set to 13K. After the clustering process there were 7720 density clusters with more than one element (multilingual clusters) and 5280 monolingual clusters. This means that 25K of the 31K language-dependent densities are mapped to a multilingual cluster. The method IPA-OVL shows a significant improvement for the American system. The decrease was now only 1.7% in comparison to the language-dependent case.

4. SUMMARY AND CONCLUSION

In this paper we demonstrated the usefulness and feasibility of the multilingual approach. First, a telephone-based multilingual speech recognition system was built for 6 languages. The language-dependent phonetic models can be used for a vocabulary independent recognition tasks. Second, we developed and compared three different methods to create multilingual phone models. The best result was achieved with the method IPA-OVL which exploits the acoustic-phonetic similarities in an optimal way. However, this method works on the density level rather than on a complete phone level. Hence, it is important to consider the language-dependent properties of the phones even if they belong to the same IPA-based phone. The main advantage of the data-driven methods are the higher recognition rate and the fact that the final number of parameters can be adjusted during clustering. In all our investigations we used only context-independent models. Now it would be interesting to know how these methods would work with context-dependent models. Further, more languages of other language families should be integrated in this multilingual approach.

5. REFERENCES

[1] O. Andersen, P. Dalsgaard, W. Barry: *Data-Driven Identification of Poly- and Mono-phonemes for four European Languages. Proc. Eurospeech 1993*, 759–762, Berlin, 1993.

[2] O. Andersen, P. Dalsgaard: *Language-Identification Based on Cross-Language Acoustic Models and Optimised Information Combination. Proc. Eurospeech 1997*, 67–70, Rhodos, 1997.

[3] K.M. Berkling: *Automatic Language Identification with Sequences of Language-Independent Phoneme Clusters.* Oregon Graduate Institute of Science & Technology, 1996.

[4] U. Bub, J. Köhler, B. Imperl: *In-Service Adaptation of Multilingual Hidden-Markov-Models. Proc. ICASSP 1997*, 1451–1454, München, 1997.

[5] P. Bonaventura, F. Gallocchio und G. Micca: *Multilingual Speech Recognition for Flexible Vocabularies. Proc. Eurospeech 1997*, 355–358, Rhodos, 1997.

[6] G.L. Campbell: *Concise Compendium of the World's Languages.* Routledge, New York, 1995.

[7] C. Corredor-Ardoy, J. Gauvin, M. Adda-Decker, L. Lamel: *Language Identification with Language-Independent Acoustic Models. Proc. Eurospeech 1997*, 55–58, Rhodos, 1997.

[8] M. Falkhausen, H. Reininger, D. Wolf: *Calculation of Distance Measures Between Hidden Markov Models. Proc. Eurospeech 1995*, 1487–1490, Madrid, 1995.

[9] J.T. Foote, H.F. Silverman: *A Model Distance Measure for Talker Clustering and Identification. Proc. ICASSP 1994*, 317–320, Adelaide, 1994.

[10] J. Glass, et al.: *Multilingual Spoken Language Understanding in the MIT VOYAGER System.* Speech Communication, vol. 17, 1–18, 1995.

[11] A. Hauenstein, E. Marschall: *Methods for Improved Speech Recognition Over the Telephone Lines. Proc. ICASSP 1995*, 425–428, Detroit, 1995.

[12] J.L. Hieronymus: *ASCII Phonetic Symbols for the World's Languages: Worldbet.* Bell Labs Technical Memorandum, 1993.

[13] International Phonetic Association: *The International Phonetic Association (revised to 1993) – IPA chart. Journal of the International Phonetic Association*, vol. 1, Nr. 23, 1993.

[14] B.H. Juang, L.R. Rabiner: *A probabilistic distance measure for hidden Markov models.* Bell Syst. Tech. J., vol. 64, Nr. 2, 391–408, 1985.

[15] J. Köhler: *Multi-Lingual Phoneme Recogntion Exploiting Acoustic-Phonetic Similarities of Sounds. Proc. ICSLP 1996*, 2195–2198, Philadelphia, 1996.

[16] J. Köhler, : *Language Adaptation of Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks. Proc. ICASSP 1998*, 417–420, Seattle, 1998.

[17] P. Ladefoged, I. Maddieson: *The Sounds of the World's Languages.* Blackwell Publishers, Oxford, 1995.

[18] L.F. Lamel, J.L. Gauvain: *Cross-Lingual Experiments with Phone Recognition. Proc. ICASSP 1993*, 507–501, 1993.

[19] L.F. Lamel, M. Adda-Decker, J.L. Gauvin: *Issues in Large Vocabulary, Multilingual Speech Recognition. Proc. Eurospeech 1995*, 185–188, Madrid, 1995.

[20] Y.K. Muthusamy, A. Cole, B.T. Oshika: *The OGI Multi-Language Telephone Speech Corpus. Proc. ICSLP 1992*, 895–898, Banff, 1992.

[21] A. Sankar, F. Beaufays, V. Digalakis: *Training Data Clustering For Improved Speech Recognition. Proc. Eurospeech 1995*, 503–506, Madrid, 1995.

[22] T. Schultz, A. Waibel: *Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets. Proc. Eurospeech 1997*, 371–374, Rhodos, 1997.